软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(2):784−798 [doi: 10.13328/j.cnki.jos.007465] [CSTR: 32375.14.jos.007465]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

结合多模态信息的定制化评论生成*

强敏杰, 王中卿, 李寿山, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 王中卿, E-mail: wangzq@suda.edu.cn

摘　要: 随着商家评论网站的快速发展, 网站上的内容越来越多, 用户难以在短时间内获取到有价值的评论. 引入了一项名为“多模态定制化评论生成”的新任务. 该任务旨在为特定用户生成他们尚未评价的产品的定制化评论, 这有助于用户对特定产品提供宝贵的意见. 为了实现这一目标, 探索了一种基于预训练语言模型的多模态评论生成框架. 具体而言, 采用了一种多模态预训练语言模型. 该模型接受产品图片和用户偏好作为输入. 之后对视觉和文本特征进行融合, 从而生成定制化评论. 实验结果表明, 该模型在生成高质量的定制化评论方面具有显著效果.

关键词: 多模态信息; 评论生成; 预训练模型; 注意力机制

中图法分类号: TP18

中文引用格式: 强敏杰, 王中卿, 李寿山, 周国栋. 结合多模态信息的定制化评论生成. 软件学报, 2026, 37(2): 784–798. http://www.jos.org.cn/1000-9825/7465.htm

英文引用格式: Qiang MJ, Wang ZQ, Li SS, Zhou GD. Customized Review Generation Integrating Multimodal Information. Ruan Jian Xue Bao/Journal of Software, 2026, 37(2): 784–798 (in Chinese). http://www.jos.org.cn/1000-9825/7465.htm

Customized Review Generation Integrating Multimodal Information

QIANG Min-Jie, WANG Zhong-Qing, LI Shou-Shan, ZHOU Guo-Dong

(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

Abstract: With the rapid development of merchant review websites, the volume of content on these websites has increased significantly, making it challenging for users to quickly find valuable reviews. This study introduces a new task, “multimodal customized review generation”. The task aims to generate customized reviews for specific users about products they have not yet reviewed, thus providing valuable insights into these products. To achieve this goal, this study explores a multimodal review generation framework based on a pre-trained language model. Specifically, a multimodal pre-trained language model is employed, which takes product images and user preferences as inputs. The visual and textual features are then fused to generate customized reviews. Experimental results demonstrate that the proposed model is effective in generating high-quality customized reviews.

Key words: multimodal information; review generation; pre-trained model; attention mechanism

随着社会的发展和生活水平的提高, 评论网站越来越受到欢迎. 这些评论网站 (如: Yelp、Amazon 等) 为客户提供了一个表达意见并根据个人体验对产品或服务进行评分的平台. 这些网站已成为潜在客户的重要信息来源, 消费者在做出购买决策之前希望找到客观的反馈. 通过提供总体评分和详细的用户评论, 这些平台为产品或服务提供了综合评价, 帮助客户做出明智的选择.

情感分析和推荐系统是分析评论网站上评论和评分的两个重要研究领域. 情感分析[1−3]旨在从评论文本中提取属性词和情感词, 为每个属性词分配一个唯一的预定义类别, 并给出该属性词的语义倾向 (例如, 积极、消极或者中立). 推荐系统[4−6]根据用户的购买历史和其他购买过目标产品的客户, 为特定产品或服务生成推荐指数. 尽管这些研究取得了显著成果, 但主要集中于分析现有的评论或购买历史, 无法生成用户在未曾评论过特定产品时可

* 基金项目: 国家自然科学基金 (62376178); 江苏省高校优势学科项目
收稿时间: 2024-10-02; 修改时间: 2025-01-16, 2025-04-16; 采用时间: 2025-05-07; jos 在线出版时间: 2025-09-24
CNKI 网络首发时间: 2025-09-25

能撰写的定制化评论. 而这种定制化评论能够根据用户的个人偏好提供特定产品的简要总结, 帮助用户快速了解产品信息.

因此, 本文提出了一个全新的任务: 多模态定制化评论生成, 其旨在针对未被评论的产品, 根据用户提供的偏好以及产品图片生成定制化评论. 通过分析定制化评论, 用户可以深入了解到特定产品围绕个人偏好的相关信息. 如图 1 所示, 展示了基于用户偏好和产品图片生成定制化评论的示例. 模型通过分析用户的偏好信息, 了解到用户喜欢猪肋排和薯条, 并对这些食物持积极的情感倾向. 同时, 模型还获取了商家的产品图片, 从而进一步了解菜品的细节, 如份量和色泽等. 结合以上信息, 模型为用户生成了一条与其个人偏好及产品图片相契合的定制化评论.

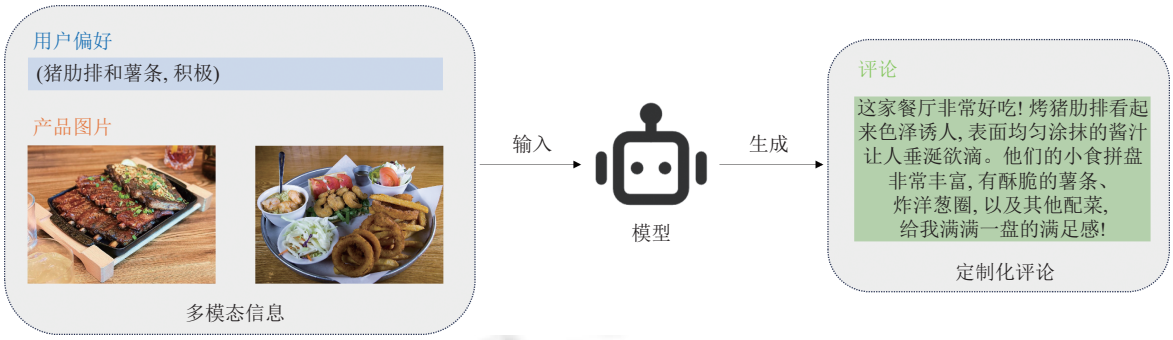

图 1 基于用户偏好和产品图片生成定制化评论的示例

生成定制化评论的一个直接方法是使用产品的图片. 然而, 仅凭图片无法完全描述产品的详细信息或有效传达关于产品的看法. 因此, 本文提出将图片与用户偏好结合使用, 以生成更全面和定制化的评论. 这里的用户偏好包括了与产品相关的属性词和情感倾向, 有助于详细阐述产品特点并反映用户的真实感受. 这种视觉和文本信息的整合使得评论更加完整和细致, 准确捕捉产品的特征和用户的意见.

近年来, 预训练模型[7,8]在多模态任务中表现出显著的优势. 通过将多种预训练模型进行结合, 能够有效地处理文本与图片数据, 这些模型利用大规模数据集进行预训练, 提前学习不同模态下的语义知识, 从而更好地捕捉文本与图像之间的复杂关系及潜在信息. 因此, 本文采用了预训练模型 T5[9]和 ViT[10]作为基础模型, 构建了一个多模态预训练语言模型. 该模型能够同时处理视觉和文本信息. 首先, 本文采用了 BLIP-2[11]模型为每张产品图片生成标题, 以弥合文本和图像之间的差距. 随后, 将产品图片、用户偏好和生成的图片标题组合, 作为模型输入送入编码器. 接下来, 将编码后的视觉和文本向量表示输入到一个基于文本引导的融合模块中, 该模块能够有效融合多种模态的信息. 最后, 模型根据融合模块输出的向量表示生成定制化评论. 实验结果表明, 本文提出的模型在指标上优于对比基准模型. 总体而言, 本文探索并证明了多模态预训练语言模型通过结合视觉与文本信息有效生成定制化评论的可行性, 为该领域未来研究奠定了基础.

本文的主要贡献如下.

- 提出了一种新的多模态定制化评论生成任务, 旨在针对未被评论的产品, 根据用户提供的偏好以及产品图片生成定制化评论.
- 提出了一个基于预训练模型的多模态评论生成框架, 其中包括多模态信息编码、图片标题生成和基于文本引导的融合模块.
- 为了评估模型的有效性, 本文进行了一系列对比与消融实验, 研究了模型各组件的性能与作用. 实验结果表明, 相比于其他最新的基准模型, 本文在 ROUGE、BLEU 和 Meteor 这 3 个评价指标上均有较大的提升.

1 相关工作

1.1 情感分析与推荐系统

早期的情感分析研究主要集中在文档级情感分类上[12−15], 这种方法涉及分析整个文档, 以判断其表达的总体

情感态度是积极的、消极的还是中立的. 尽管这些早期方法为情感分析领域奠定了基础, 它们通常忽略了文本中许多重要的情感信息. 随着研究的深入, 学者们追求更高的分析精度和更细的粒度, 这促使情感分析领域的焦点逐渐转向基于方面的情感分析 (aspect-based sentiment analysis, ABSA). ABSA 的研究起初集中于单个子任务, 如属性词提取 (aspect term extraction, ATE)[16]、属性词类别识别 (aspect category detection, ACD)[17]和属性词情感分类 (aspect sentiment classification, ASC)[18]. 随着技术的发展, 研究开始探索更复杂的任务组合, 包括同时提取属性词和情感词[19,20], 以及识别特定属性词的类别和相应的情感极性[21,22]. 最近, 端到端模型也被用于提取三元组或四元组格式的情感元素[23,24], 并在多个情感元素提取任务中取得了令人印象深刻的性能. Bao 等人[3]引入了一种新的意见树生成模型, 旨在联合检测树中的所有情感元素. 意见树可以揭示更全面、更完整的方面级情感结构. 此外, 他们采用预训练模型来集成语法和语义特征以生成意见树. 为了解决多重重叠三元组带来的困扰, Zhao 等人[23]提出了一种新颖的多重重叠三元组提取方法. 该方法通过学习多个属性词和情感词之间的协作相互作用来解码它们之间的复杂关系. Gou 等人[24]提出了多视图提示 (multi-view prompting, MvP), 它聚合了以不同顺序生成的情感元素, 利用来自不同视图的类人问题解决过程的直觉.

推荐系统是一项被广泛应用的任务, 其目的是根据用户的偏好和历史行为为用户提供定制化的推荐. 这一领域的早期研究主要集中在协同过滤技术上[25−27], 这是一种通过分析用户间相似性来预测用户可能喜欢的项目的方法. 然而, 由于协同过滤方法在处理数据稀疏性时的局限性, 导致推荐效果不佳, 因此一些研究开始探索利用用户的历史评论来缓解这一问题. 这些方法主要分为两类. 一类是历史评论方法[28,29]. 该方法通过分析用户和项目的评论内容来学习更丰富的词嵌入, 从而更好地理解用户和项目之间的关系. 另一类是目标评论方法[30−33]. 这种方法不仅分析历史评论, 而且在模拟用户与项目之间的交互时, 利用目标评论来提高推荐的准确性. 具体来说, 这种方法在训练阶段通过分析用户-项目间的互动来构建模型, 并在推理阶段使用类似 Transformer 的层来精确捕捉目标评论中的细微差别, 以此来优化推荐结果.

1.2 评论生成

近年来, 评论生成作为自然语言生成的一个重要分支, 受到了广泛关注. 该任务的目标是根据给定的输入生成与上下文一致且富有信息性的评论. 早期的工作利用用户或者产品的基本信息 (例如 ID) 来生成评论[34,35]. Tang 等人[35]使用评论以及项目 ID 作为输入, 利用门控机制来缓解长距离依赖问题, 应用基于 RNN 的解码器来生成最终的评论. Ni 等人[30]提出了 ExpansionNet. 该框架基于编码器-解码器框架, 使用产品标题和评论摘要作为模型输入, 通过基于感知的方面编码器来合并方面级信息, 并利用注意力融合模块来关注多个编码器的输出. Li 等人[31]提出了 RevGAN 模型. 该模型由 3 个组件构成: 自注意递归自编码器、条件判别器和个性化解码器, 利用项目描述和情感标签来生成定制化评论. 最近, 大语言模型的兴起在多个 NLP 任务上取得了显著的进展. Peng 等人[36]首先通过汇总用户的历史行为来构建提示输入, 然后将作为满意度指标的评分纳入提示中, 这可以进一步提高模型对用户偏好的理解以及对所生成评论的情感倾向控制. 最后, 将提示文本输入到 LLM 中生成定制化评论.

1.3 多模态融合

多模态融合通过结合来自不同模态的信息, 有效提升了模型的整体性能. 这种技术使模型能更好地理解和处理复杂数据, 从而提高决策的准确性和鲁棒性[37,38]. 在多模态情感分析场景中, Zadeh 等人[39]将多模态情感分析问题视为建模模态内和模态间动态的问题, 引入了一种新颖的模型, 称为张量融合网络. 该模型能够端到端地学习这两种动态情况. 随着 Transformer 的普及, Tsai 等人[40]提出了多模态转换器 (MulT), 以端到端的方式解决数据不对齐问题, 而无需显式对齐数据. 其模型的核心是定向成对跨模态注意力机制. 该机制关注不同时间步长上多模态序列间的交互, 并潜在地将数据流从一种模态适应到另一种模态. Huang 等人[41]引入了多模态 Transformer 在模型层面融合视-听模态, 从而缓解数据对齐失败和长距离依赖问题. 最近, Yang 等人[42]提出了自适应上下文与模态交互建模 (self-adaptive context and modal-interaction modeling, SCMM) 框架. 不再满足于简单地连接模态特征, 而是根据模态对整体贡献的不同来分别处理, 以充分挖掘模态间的交互作用.

本文提出的任务在输入和输出方面与多个相似任务有显著不同. 例如, 在情感分析中, 输入通常是已有的评

论, 输出是属性词或情感词; 在推荐系统中, 输入通常是用户或商品的 ID, 并基于历史评论推断用户偏好和商品属性, 输出是推荐结果及理由; 在评论摘要生成任务中, 输入是完整的评论文本, 输出是精简的摘要; 而在虚假评论检测任务中, 输入是已有的评论, 目标是判断评论的真实性. 相比之下, 本文的任务输入为用户偏好和产品图像, 输出是基于这些信息生成的定制化评论. 本文任务的重点在于为尚未评论的产品生成评论, 且仅使用用户偏好和产品图像作为输入, 从而确保评论的个性化和相关性. 这使得本文的任务不仅不同于传统的定制化评论生成方法 (通常依赖历史数据或用户档案), 而且避免了评论摘要生成和虚假评论检测任务中对于已有评论的依赖.

2 基于预训练模型的多模态评论生成

本文研究了一项新任务: 多模态定制化评论生成. 该任务根据用户偏好和相关图像为产品生成定制化评论. 形式上, 该任务的输入是一个元组 $\{I,T\}$, 其中, $I=\{I_1,I_2,\ldots,I_n\}$ 表示产品图像, $T=\{T_1,T_2,\ldots,T_n\}$ 表示描述用户偏好的输入词. 本文的模型输出为一个定制化评论 C. 该评论为产品提供了详细且定制化的描述. 这一任务具有较高的挑战性, 因为评论需要根据特定产品和用户偏好进行定制, 这使得任务变得复杂, 要求对语言和视觉信息有深入的理解.

如图 2 所示, 本文提出一种基于预训练语言模型的多模态评论生成框架, 以应对上述挑战. 为了在文本和图像模态之间建立无缝连接, 框架引入一个图像标题生成模块. 该模块为图片生成描述性且与上下文相关的标题. 随后, 利用文本编码器将输入文本转化为丰富的文本特征表示, 并采用图像编码器将相应图片编码为视觉特征表示. 接着, 利用基于文本引导的融合模块来整合文本和视觉特征表示. 最终, 融合后的向量表示被输入到解码器中, 用于生成定制化评论. 为了实现良好的参数初始化, 本文采用 T5 和 ViT 作为基础模型, 两者均为基于 Transformer 结构的预训练模型, 已从广泛的数据集中获取了丰富的语义知识. 接下来将详细描述各个模块的具体实现.

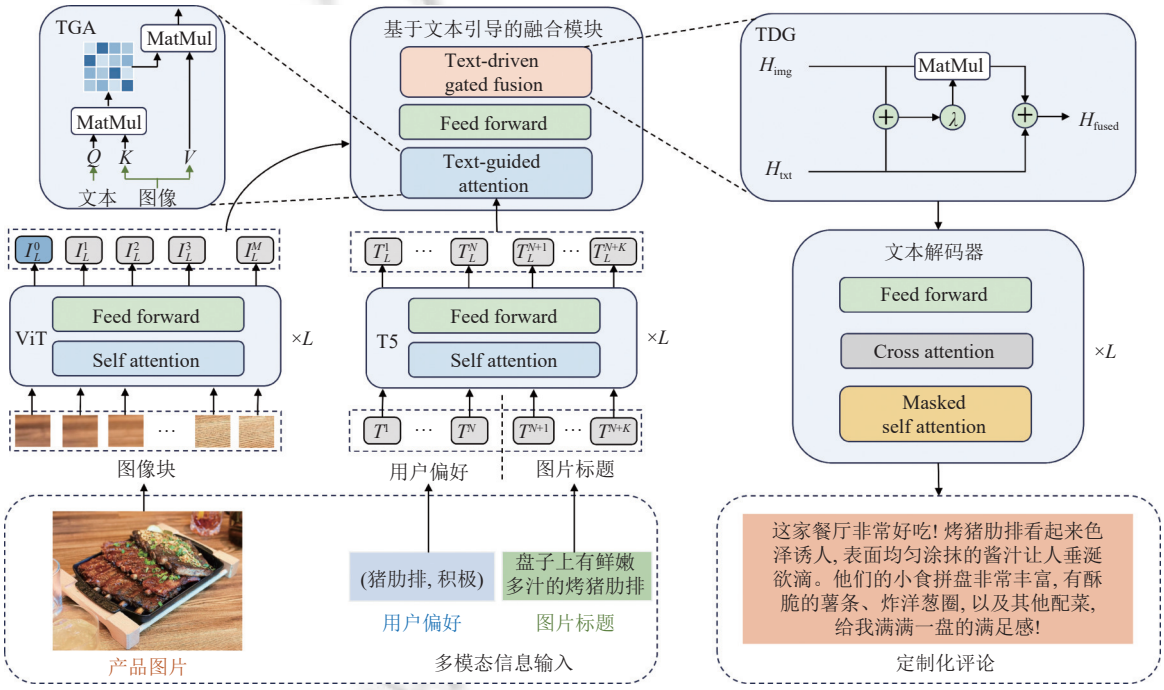

图 2 基于预训练语言模型的多模态评论生成框架

2.1 文本编码器

文本编码器的输入为一段文本 T. 该文本由用户偏好和产品图片标题组成. 用户偏好包括属性词和情感极性,

属性词用于描述用户对特定属性的偏好, 情感极性则表示用户对这些属性的情感倾向. 本文通过对文本 T 进行编码, 生成文本特征向量.

具体地, 首先对输入文本 T 进行分词处理, 并输入到词嵌入层中, 获得词嵌入向量 W. 然后, 将词嵌入向量 W 与位置向量 E_{pos} 相加, 得到输入特征 T_0:

$$T_0 = W + E_{\text{pos}},\ \{T_0, W, E_{\text{pos}}\} \in \mathbb{R}^{N\times D} \tag{1}$$

其中, N 表示序列长度, D 表示特征维度. 接下来, 输入特征 T_0 输入到文本编码器中. 文本编码器由 L 层相同结构的模块堆叠而成, 每一层包含一个多头自注意力 (self-attention, *MSA*) 子层和一个前馈网络 (feed-forward network, *FFN*) 子层, 每一层的计算过程如下:

$$T'_\ell = LN(MSA(T_{\ell-1}) + T_{\ell-1}),\ T'_\ell \in \mathbb{R}^{N\times D} \tag{2}$$

$$T_\ell = LN(FFN(T'_\ell) + T'_\ell),\ T_\ell \in \mathbb{R}^{N\times D} \tag{3}$$

其中, T_ℓ 表示第 ℓ 层编码器的隐藏状态, $LN(\cdot)$ 表示层归一化 (layer normalization, *LN*) 操作.

2.2 图片编码器

图片编码器的输入是一组产品图片 I, 这些图片被视为一个整体并依次拼接在一起, 以描述产品的各项细节, 如颜色、质地等. 本文将拼接后的产品图片输入到图像编码器中, 生成视觉特征向量.

具体地, 为了使用 ViT 编码图片, 首先将图片分割成一系列扁平化的二维块序列. 然后, 在块序列的开头添加一个特殊标记 [CLS], 形成输入序列 X. 接着, 将输入序列 X 和位置向量 E_{pos} 相加, 得到输入特征 I_0:

$$I_0 = X + E_{\text{pos}},\ \{I_0, X, E_{\text{pos}}\} \in \mathbb{R}^{(M+1)\times D} \tag{4}$$

其中, M 表示序列长度. 最后, 将输入特征 I_0 输入到 ViT 中, 获得特殊标记 [CLS] 的视觉表示 H_{img}. 虽然 ViT 和 Transformer 编码器结构相似, 但需注意层归一化的位置有所不同, 每一层的计算过程如下:

$$I'_\ell = MSA(LN(I_{\ell-1})) + I_{\ell-1},\ I'_\ell \in \mathbb{R}^{(M+1)\times D} \tag{5}$$

$$I_\ell = MLP(LN(I'_\ell)) + I'_\ell,\ I_\ell \in \mathbb{R}^{(M+1)\times D} \tag{6}$$

$$H_{\text{img}} = LN(I_L^0) \tag{7}$$

其中, I_ℓ 表示第 ℓ 层编码器的隐藏状态, I_L^0 表示来自第 L 层的特殊标记 [CLS] 输出, $MLP(\cdot)$ 是一个全连接层.

2.3 图像标题生成

现有的多模态研究表明, 与经典的单模态任务相比, 加入视觉模态并不总能带来显著的改进, 甚至可能产生负面影响. Zhu 等人[43]认为, 图像中可能存在噪声, 而输入文本已经包含了生成目标文本所需的足够信息. 因此推测当前的方法并未充分挖掘视觉模态中有意义信息的潜力. 为了解决这一问题, 本文提出利用图像标题作为桥梁, 连接图像与文本. 具体来说, 在生成图像的视觉表示的同时, 将其转换为自然语言形式, 从而更有效地利用视觉信息.

如后文图 3 所示, 本文采用 BLIP-2 模型来生成图像标题. 这是一个在多个视觉-语言任务上展现出卓越性能的大型多模态语言模型. 其中, 图像编码器使用的是 CLIP 的 ViT, 而基础语言模型建立在基于编码器-解码器的 FlanT5 模型[44]之上并在 COCO 数据集[45]上对其进行微调, 让其适用于图像标题生成任务. 具体来说, 给定多模态数据 $\{I, T\}$, 本文使用 BLIP-2 模型为产品图像 I 生成相应的标题. 然后, 将图像标题与输入文本 T 进行拼接, 使得输入文本的最终格式为“⟨属性词⟩; ⟨情感极性⟩; ⟨图像标题⟩”.

2.4 基于文本引导的融合模块

在获得文本表示 H_{txt} (= T_L) 和视觉表示 H_{img} 的 [CLS] 标记后, 本文设计了一个基于文本引导的融合模块, 以实现多模态融合. 该模块包括 3 个子层: 文本引导的注意力 (text-guided attention, TGA) 融合子层、前馈网络子层和文本驱动的门控 (text-driven gated, TDG) 融合子层, 如图 2 所示.

该融合模块的优势在于, TGA 子层能够根据一组属性词和情感极性, 学习每个图像对象的不同注意力分数,

从而聚焦于图像的特定部分. 例如, 针对属性词“猪肋排”和情感极性“积极”, TGA 子层能够帮助模型关注图像中包含猪肋排的部分, 以生成具体的正面评价. 在 TDG 子层中, 通过门控因子 λ 控制保留多少视觉信息, 从而过滤掉冗余和噪声. 接下来详细介绍融合模块中的各个子层.

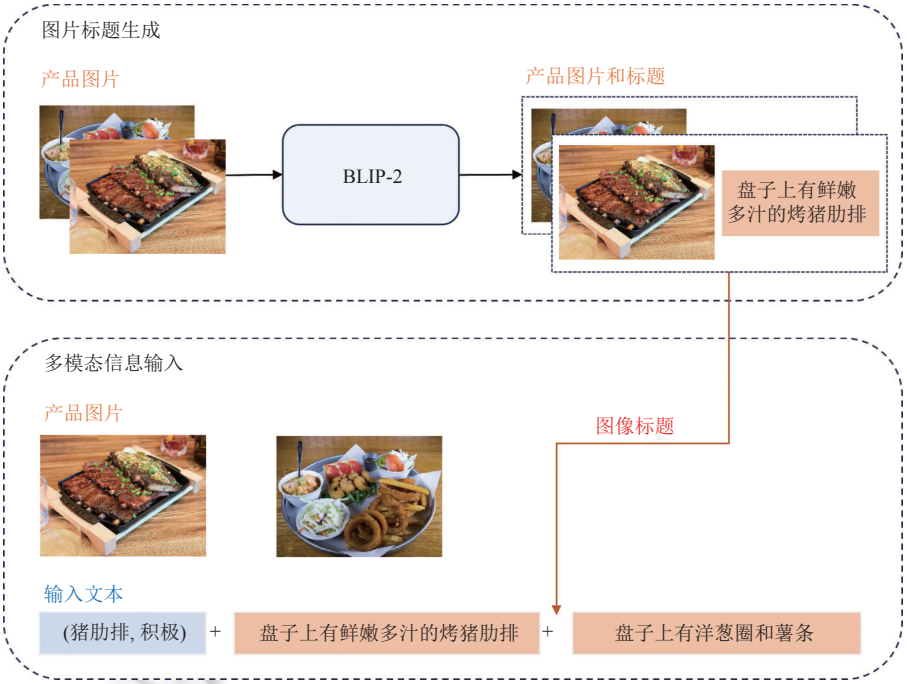

图 3 图像标题生成

● 文本引导的注意力融合子层 (TGA). 由于视觉表示 H_{img} 的序列长度为 1, 首先使用广播机制将其扩展到与文本序列一致的长度, 以避免维度不匹配问题. 接着, 将文本表示投影到 Query 空间, 将视觉表示投影到 Key 和 Value 空间, 应用文本引导的注意力机制获取文本引导的视觉特征 Z_{img}:

$$Q = H_{\text{txt}}W^{Q},\ [K;V] = H_{\text{img}}[W^{K};W^{V}] \tag{8}$$

$$Z_{\text{img}} = TGA(Q,K,V) \tag{9}$$

$$H_{\text{txt}} = LN(H_{\text{txt}} + Z_{\text{img}}) \tag{10}$$

其中, $\{W^{Q},W^{K},W^{V}\}$ 是训练参数. 随后, H_{txt} 将通过前馈网络子层:

$$H_{\text{txt}} = LN(H_{\text{txt}} + FFN(H_{\text{txt}})) \tag{11}$$

● 文本驱动的门控融合子层 (TDG). 利用门控机制整合来自不同模态的表示. 首先, 通过门控因子 λ 衡量模态间的冗余与视觉信息中的噪声量:

$$\lambda = Sigmoid(W^{T}H_{\text{txt}} + W^{I}H_{\text{img}}) \tag{12}$$

其中, W^{T} 和 W^{I} 是训练参数. 接着, 将门控后的视觉信息融入文本特征中, 得到融合特征 H_{fused}:

$$H_{\text{fused}} = H_{\text{txt}} + \lambda \cdot H_{\text{img}} \tag{13}$$

值得注意的是, 本文的方法完全保留了原始文本特征, 这与传统的门控机制有所不同. 其背后的直觉是, 文本信息相比于视觉信息更为重要, 因此本文的设计旨在最大程度保留文本信息, 同时结合视觉模态的基本信息.

2.5 定制化评论生成

在多模态表示融合完成后, 解码器逐标记地生成输出序列. T5 解码器与文本编码器结构类似, 均由 L 层相同

的子层组成, 但在每层中增加了一个多头交叉注意力子层 (multi-headed cross-attention, MCA). 在每个解码时间步 t, 第 ℓ 层解码器的隐藏状态 O_ℓ 通过以下公式计算:

$$O'_\ell = LN(MSA(O_{\ell-1})) + O_{\ell-1} \tag{14}$$

$$O''_\ell = LN(MSA(O'_\ell, H_{\text{fused}}) + O'_\ell) \tag{15}$$

$$O_\ell = LN(FFN(O''_\ell) + O''_\ell) \tag{16}$$

生成的输出序列以结束标记“</s>”结尾. 最终, 整个序列的条件概率 $p(y|I,T)$ 通过每一步的概率 $p(y_t|y_{<t},I,T;\theta)$ 计算得到:

$$p(y|I,T) = \prod_{t=1}^{|y|} p(y_t|y_{<t},I,T;\theta) \tag{17}$$

$$p(y_t|y_{<t},I,T;\theta) = \sigma(W^o O_{L,t} + b^o) \tag{18}$$

其中, $\{W^o, b^o\}$ 是可训练参数, $\sigma(\cdot)$ 是 *Softmax* 函数, $y_{<t} = y_1 \ldots y_{t-1}$ 和 $p(y_t|y_{<t},I,T;\theta)$ 是通过 *Softmax* 归一化的目标词汇表上的概率.

3 实 验

本节中进行了大量的实验以评估本文提出模型的性能. 此外, 还提供了各种分析和讨论, 以展示本文模型的有效性.

3.1 数据与设置

本文构建了一个新的数据集, 用于多模态定制化评论生成任务. 该数据集基于 Gest[46]. Gest 数据集主要来自 Google Local (即谷歌地图) 的餐馆评论和图片. 数据收集过程采用了广度优先搜索算法, 目标是从餐馆页面抓取评论和相关的用户生成的图片. 在数据收集后, 移除评论长度小于 5 个单词的记录以及超过 10 张图片的评论, 最后划分成了两个子集 Gest-s1 和 Gest-s2. 其中, Gest-s2 经过了人工标注, 确保每个图像-文本对能够准确地反映图像的内容, 因此本文选择 Gest-s2 作为基础数据集. 本文数据集中的每条评论文本至少拥有一张对应的图像. 随后, 使用情感分析模型[3]提取情感元素, 并将其作为用户偏好的输入. 数据集的统计信息详见表 1. 实验中, 本文随机选取了 4 000 个样本作为训练集, 500 个样本作为开发集, 其余样本用作测试集.

表 1 数据集的统计数据

类别	数量
每个样本的平均图片数	2.29
每个样本的平均输入字数	4.99
定制化评论的长度	47.62
数据样本总数	5 000

本文使用 T5-Base 和 ViT 作为基座模型, 并使用 Adam[47]作为优化器在动量 $\beta = 1$ 的条件下去微调超参数. 本文模型的学习率被设置为 0.000 1, Batch 大小设置为 4, 共进行 20 轮训练. 为了生成更高质量的评论, 本文使用了束搜索 (beam search) 技术, Beam 大小设置为 5 并限制了重复 n-grams 大小为 3[48]. 所有实验均在 NVIDIA Tesla V100 16 GB 显卡上完成.

本文在实验中采用了 3 种主流的自动评估指标, 即 ROUGE[49]、BLEU[50]和 Meteor[51], 以全面评估模型生成的文本质量. 这些指标分别从不同的角度对生成内容进行量化分析: ROUGE 主要关注生成文本与参考文本之间的重叠情况, BLEU 则衡量生成文本与参考文本在词序上的相似度, 而 Meteor 则综合考虑了词形变化和语义匹配的因素. 通过这些指标, 对模型生成的评论质量进行了细致分析, 验证了本文模型在实际应用中的有效性.

3.2 基准模型

为了评估在多模态定制化评论生成任务上的表现, 本文选取了几种基准模型在数据集上进行了微调并对比,

其中包含 Opword、BART[52]、T5、LLaMA[53]、GPT-4o[54]、ViT-GPT2、GIT[55]、Pix2Struct[56]、BLIP-2[11]、LLaVA-1.5[57]、Selective-Attn[58]、VLP-MABSA[59]和 AoM[60]. 下面是这些基准模型的简要介绍.

● Opword: 直接将给定的输入文本作为定制化评论.

● BART: 一个结合了双向编码器和自回归解码器的序列到序列模型. 通过将输入数据进行噪声化处理并恢复, BART 在文本生成任务中表现出色, 尤其在摘要生成、翻译和文本生成等方面.

● T5: Google 提出的一个将所有 NLP 任务统一为文本到文本格式的模型. T5 的核心思想是将输入和输出都作为文本进行处理, 使得它能够广泛适用于翻译、摘要、分类和问答等多种任务, 体现了极大的灵活性和强大的通用性.

● LLaMA: Meta 公司开发的大型语言模型, 旨在提供更高效的计算和内存利用率. LLaMA 模型在多语言处理、自然语言生成等任务中表现优异, 适合在有限资源下执行复杂的自然语言处理任务.

● GPT-4o: 基于 OpenAI 的 GPT 架构开发的对话生成模型, 通过在大量的文本数据和人类对话数据上进行训练, 它能够生成连贯且上下文相关的对话. GPT-4o 广泛应用于聊天机器人、客户服务、教育和娱乐等领域.

● ViT-GPT2: 一个多模态模型, 由一个图像编码器 (ViT) 和文本解码器 (GPT2) 组成, 能够基于给定的图像生成自然语言文本.

● GIT: 一种生成式图像到文本的 Transformer, 用于统一图像/视频字幕和问答等视觉语言任务. 它简化了现有的多模态编码器/解码器结构, 并在单一语言建模任务下只采用了一个图像编码器和一个文本解码器. 同时, 通过扩大预训练数据和模型尺寸, 提高了模型性能.

● Pix2Struct: 一种专为纯视觉语言理解设计的预训练图像到文本模型, 适用于对包含视觉情境语言的任务进行微调. 值得注意的是, Pix2Struct 的预训练任务是通过学习将屏幕截图解析为简化的 HTML 来完成的.

● BLIP-2: 一个大型多模态语言模型, 能够利用预训练的冻结图像编码器和大规模语言模型来指导视觉语言的预训练. BLIP-2 通过 Querying Transformer 来实现跨模态的对齐, 并在多个视觉语言任务上取得了优异的性能.

● LLaVA-1.5: 一种多模态大模型, 它通过将视觉编码器与大语言模型结合, 能够理解图像内容并进行基于图像的对话和推理.

● Selective-Attn: 一种利用选择性注意机制增强 ViT 特征融合表示的方法, 旨在增强文本和视觉特征对模型性能的贡献.

● VLP-MABSA: 一种针对 MABSA 的任务特定视觉语言预训练框架. 该框架是所有预训练和下游任务的统一多模态编码器-解码器架构. 该框架分别从语言、视觉和多模态设计了 3 种类型的特定任务预训练任务.

● AoM: 设计了一个属性词感知注意模块来同时选择与属性词语义相关的文本标记和图像块. 为了准确聚合情感信息, 它将情感嵌入引入 AoM 中, 并使用图卷积网络对视觉-文本和文本-文本交互进行建模.

3.3 实验结果与分析

如表 2 所示, 传统单模态方法 (如 T5 和 BART) 在 ROUGE-L 和 BLEU 等指标上表现出较为显著的优势, 其原因在于语言模型擅长处理文本特定任务, 能够高效建模上下文语义. 然而, 由于其输入仅限于文本, 难以充分捕捉多模态情境中图像信息的复杂性. 这些方法的输入词汇过于简单, 信息量有限, 无法充分反映评论的具体内容和情感色彩. 此外, 以语言模型为主的 GPT-4o 尽管因其大规模预训练展现了良好能力, 但在本文任务中的表现略逊于 T5 和 BART, 其中可能的原因是 GPT-4o 缺乏对该任务的特定优化, 生成结果的可控性较差.

与此同时, 预训练语言模型 (如 T5 和 BART) 在处理复杂信息时的性能通常优于仅依赖图像输入的方法 (如 ViT-GPT2 和 Pix2Struct). 这表明, 文本输入相比图像能够提供更直接且详细的信息, 视觉模型由于仅依赖图像输入, 表现出显著的局限性, 虽然视觉模型能够捕获图像的底层特征, 但难以生成细粒度的文本描述, 导致 ROUGE 和 BLEU 指标普遍低于语言模型.

值得注意的是, 多模态方法 (如 BLIP-2 和 LLaVA) 在所有单模态方法中表现最佳. 这一结果表明, 多模态信息在生成任务中的互补性, 文本提供了结构化和明确的语义信息, 图像则补充了丰富的视觉线索, 从而显著提升模型

的整体表现.

此外, 本文提出的模型在所有基准模型中表现出显著优势 (统计显著性 $p<0.05$). 这一结果不仅验证了模型的有效性, 同时也突显了多模态评论生成框架的优势. 该框架通过融合图像标题生成和文本引导的方式, 充分利用多模态信息, 在复杂的实际应用场景中实现了更高的性能和更好的用户体验.

表 2 对比实验结果

模型		ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-4	Meteor
语言模型	Opword	7.81	3.97	7.81	0.03	0.03	2.06
	T5	29.34	6.85	18.78	24.76	12.16	16.95
	BART	29.18	6.59	19.09	24.83	12.23	16.74
	LLaMA	26.97	6.48	19.02	21.81	11.28	15.02
	GPT-4o	27.57	6.02	18.23	22.70	11.40	15.63
视觉模型	ViT-GPT2	23.84	3.20	15.41	16.86	9.57	15.98
	GIT	24.33	3.15	16.60	21.05	10.15	14.11
	Pix2Struct	25.31	2.38	16.14	21.78	10.07	14.01
多模态模型	BLIP-2	30.25	7.58	20.61	21.02	9.50	15.25
	LLaVA-1.5	30.57	8.10	20.49	22.65	10.25	16.12
	Selective-Attn	30.23	6.79	19.00	25.23	12.14	17.44
	VLP-MABSA	29.81	8.76	22.73	23.25	11.94	17.40
	AoM	30.40	**9.14**	**23.71**	21.70	10.67	16.79
Ours		**31.76**	7.65	19.99	**26.90**	**12.73**	**18.57**

3.4 人工评价

本节将从 5 个维度对本文模型与基准模型生成的定制化评论进行全面的人工评价. (1) 流畅度: 评估评论中的语法准确性、表达流畅性以及整体语言的可读性, 衡量生成文本的文法和风格的一致性. (2) 信息量: 衡量定制化评论与参考评论之间关键信息的重合程度, 评估生成内容的准确性和信息完整性. (3) 自然度: 评估生成的定制化评论在语言风格上与人类写作的评论的相似度, 判断其是否具有人类表达的特征. (4) 情感相关度: 评估生成评论在情感表达上与参考评论的契合度, 分析情感传递的一致性. (5) 多样性: 衡量定制化评论在词汇选择、句子结构以及表述方式上的变换能力, 分析模型生成的评论内容是否重复、缺乏创意性. 为了确保评估的客观性与可靠性, 我们邀请了 3 名人工标注员对测试集中生成的定制化评论进行评分, 每个维度的评分范围为 1–5 分. 人工评价结果如表 3 所示.

表 3 人工评价结果

模型	流畅度	信息量	自然度	情感相关度	多样性
T5	3.78	3.90	3.94	3.45	2.83
GPT-4o	4.28	3.67	4.17	4.03	3.97
LLaMA	4.02	3.54	3.92	3.89	3.51
Pix2Struct	3.02	2.17	3.21	2.50	2.12
BLIP-2	4.18	4.26	4.02	4.14	3.83
Ours	**4.68**	**4.76**	**4.49**	**4.55**	**4.23**

如表 3 所示, 在评估定制化评论的生成质量时, 本文提出的模型在所有评估维度上均优于其他基准模型. 特别是在情感相关度方面, 本文模型以 4.55 的分数, 展现了显著的优越性. 这一结果表明, 本文模型在捕捉和表达定制评论中的情感细节方面具有较高的准确率和敏感性. 此外, 本文模型在多样性维度上的表现也优于其他基准模型. 这一结果表明, 本文模型在生成评论时不仅能够保持上下文的关联性, 还能够展现出丰富的词汇多样性和灵活的表达方式. 这一能力尤其重要, 因为用户评论的多样性直接关系到模型的应用前景. 在流畅度、信息量以及自然度方面, 本文模型同样表现出较高的得分, 这些都进一步证明了本文融合策略在理解和生成复杂内容方面的有效性. 例如, 本文模型在流畅度上的评分为 4.68, 领先单模态模型中的最高分 (GPT-4o 模型的 4.28), 这体现了它在语言

生成的连贯性和语法正确性方面的能力. 而在信息量和自然度上, 多模态模型同样保持了较高的标准, 分别以4.76 和 4.49 的得分领先于其他模型. 总之, 本文模型在多个方面超越了所有基准模型, 并展示了其生成内容不仅更加丰富, 而且情感表达更为精确的能力.

3.5 多模态融合策略的影响

本节将深入探讨不同多模态融合策略在生成定制化评论方面的效果. 多模态融合策略的选择对于最终模型的性能有着决定性影响, 本文评估了以下几种策略: “仅文本”策略仅使用文本数据作为输入, 不考虑任何视觉信息; “仅图片”策略仅依赖于图像数据, 不利用任何文本信息; “相加策略”通过直接相加文本和图像的表示向量进行简单的向量融合; “拼接策略”通过拼接文本和图像的表示向量来进行向量融合; “门控机制”代表使用门控机制根据模型的当前状态动态调整文本和图像信息的贡献度; “注意力机制”表示使用注意力机制融合两个向量, 允许模型根据上下文动态地聚焦于文本或图像信息的关键部分.

如表 4 所示, 实验结果表明简单的融合方法虽然提供了基本的多模态交互, 但其性能通常不如更复杂的融合策略. 这表明单纯的向量操作可能无法充分捕捉和利用文本与图像之间的复杂相互作用.

表 4 多模态融合策略的影响

方法	ROUGE-1	BLEU-1	Meteor
仅文本	29.34	24.76	16.95
仅图片	26.59	22.24	14.24
相加	29.81	24.94	16.96
拼接	29.76	25.23	17.14
门控机制	30.33	25.32	17.28
注意力机制	30.74	26.18	17.83
Ours	**31.76**	**26.90**	**18.57**

进一步地, 门控和注意力机制展现出了显著的性能优势. 这些高级融合策略能够更精确地建模不同模态间的相互依赖, 有效提升信息整合的质量和效率. 门控机制通过调节信息流动优化了信息的选择性整合, 而注意力机制则通过加权重点信息提高了模型对关键特征的敏感度和反应能力.

最重要的是, 本文提出的多模态融合策略综合了门控和注意力机制的优点, 在所有比较的策略中实现了最高的性能指标. 这一结果不仅验证了采用复合策略的有效性, 也强调了全面利用和整合多模态数据中丰富信息和关系的重要性.

3.6 文本输入的影响

表 5 中的实验结果揭示了不同输入提示策略对评论生成性能的显著影响. 本文设计了以下几种场景: “仅图片”表示模型仅使用图像信息, 不接收任何文本输入; “属性词”表示模型除了图像外, 还融合了用户偏好中的属性词; “情感倾向”表示模型除了图像外, 还融合了用户偏好中的情感极性; “图片标题”表示模型利用了图像标题, 以提供更具体的上下文信息.

表 5 文本输入的影响

方法	ROUGE-1	BLEU-1	Meteor
仅图片	26.59	22.24	14.24
+属性词	30.73	26.15	17.79
+情感倾向	27.35	22.95	14.83
+图片标题	28.39	23.86	15.44
+All (Ours)	**31.76**	**26.90**	**18.57**

实验结果表明, 仅依靠图像的模型在所有指标上的表现均较低. 这反映出图像独立使用时, 虽然能提供视觉信息, 但缺乏足够的语义深度和具体上下文, 不足以支持生成有针对性和丰富性的评论.

相比之下, 结合图像和特定文本提示的模型显示出了更高的性能. 特别是引入属性词带来的性能提升最为显著. 这表明正确的语义提示可以极大地丰富模型的理解能力, 使其能够生成更加精确和深入的评论. 在采用情感极性和图像标题作为输入时, 模型性能增幅不及属性词显著. 这可能是因为情感极性和图像标题虽然提供了情感色彩和描述性信息, 但在指导模型分析方面不如属性词有效.

值得注意的是, 综合所有文本提示的策略在所有评价指标上均达到了最佳性能. 这一结果突出了综合多种文本输入策略的重要性, 证明了在多模态融合过程中, 多种语义层面的相互作用是提升评论质量和相关性的关键.

3.7 图片和文本的影响

图 4 展示了通过消融研究评估图像和文本输入对性能的贡献. 本文设置了 3 种不同的输入场景: “全匹配”表示图像和所有相关文本输入都是准确配对的; “仅匹配”表示仅当前类别正确配对, 其余类别未配对; “仅不匹配”表示当前类别未配对, 而其余类别保持正确配对. 本文使用 T5 模型作为性能基准参考.

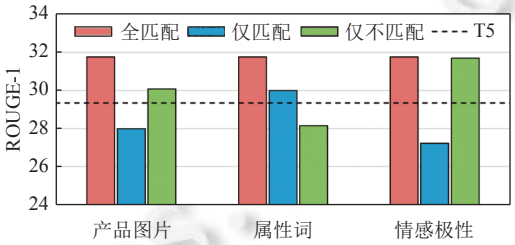

图 4 图片和文本的影响

从实验结果中可以直观地看到, 情感极性的贡献最小. 实验数据显示, 即使在仅情感极性配对的场景下, 性能也没有显著提升, 这可能是因为它没有提供实质性的有效信息.

在仅图像未配对的情况下, 模型性能显著下降, 这突显了视觉信息在维持模型性能中的核心作用. 这一结果表明, 视觉内容的正确解读对于生成准确的评论至关重要. 在仅属性词配对的情况下, 性能仍然超过基准, 表明视觉信息不仅提供了多模态信息, 还起到了正则化项的作用. 此外, 当属性词未配对时, 模型性能下降到基准水平以下, 暗示了其在模型中的关键作用.

3.8 图片数量的影响

图 5 展示了图像数量对评论生成性能的影响. 首先, 当仅依赖文本输入时, 模型的性能相对较低, 这表明纯文本方法在生成令人满意的评论方面存在局限性. 这种低性能反映出文本信息在没有视觉上下文支持的情况下, 可能无法充分表达评论所需的细节.

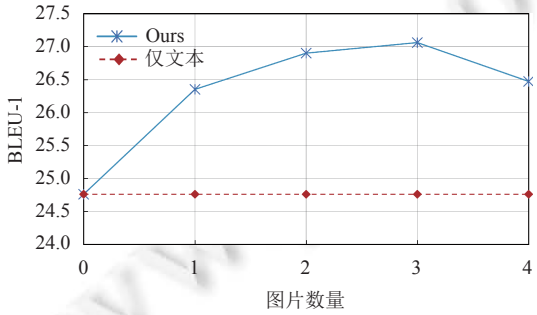

图 5 图片数量的影响

随着图像数量的增加, 性能呈现出明显的上升趋势, 特别是当图像数量从 1 增加到 3 时, 性能提升最为显著. 这种趋势表明图像为模型提供了额外的上下文信息, 有助于更精确和具体地理解和生成评论内容. 图像通过增加与评论相关的视觉细节和环境信息, 使得生成的评论更为丰富和针对性, 从而提升了评论的质量和相关性. 然而,

值得注意的是, 当图像数量达到一定程度后, 性能增长开始放缓, 并在图像数量为 5 时略有下降. 这表明在一定数量之后, 额外的图像可能不再带来性能的线性提升, 甚至可能引起信息过载, 从而影响模型的处理效率和输出质量.

总之, 图像的数量对生成的定制评论质量起着关键作用. 实验结果强调了适当数量的图像可以显著提升生成评论的总体质量和有效性, 但同时也揭示了过多图像可能导致的效益递减.

3.9 实例分析

本节展示了一个来自测试数据的样本, 用于比较本文提出的模型与其他基准模型生成的定制评论的质量. 图 6 清晰地揭示了不同模型在评论生成上的表现差异.

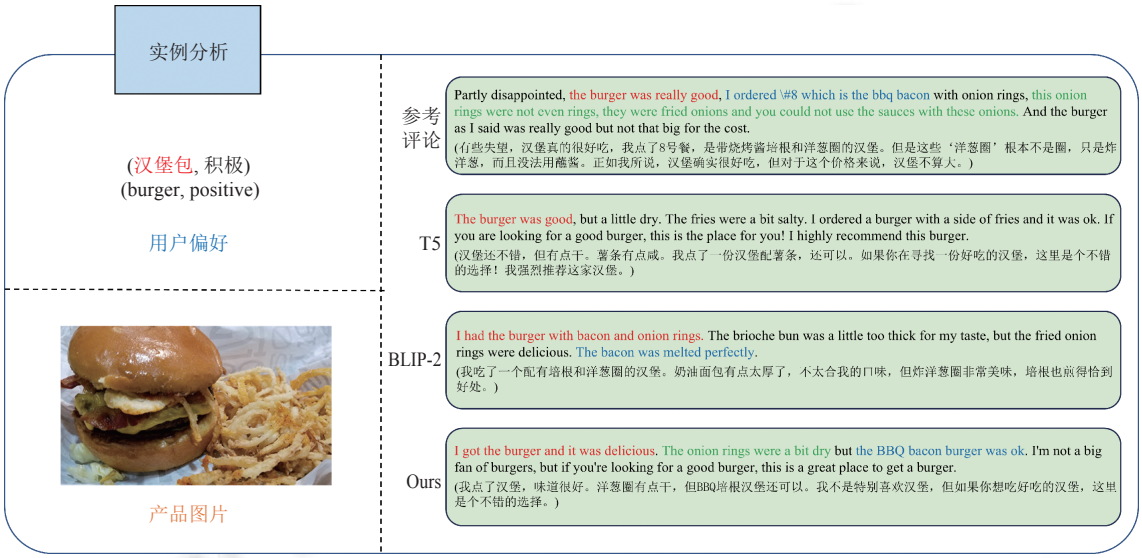

图 6 本文模型与基线模型的实例分析图

首先, 单模态预训练模型 (例如 T5) 在生成的评论中提供的信息量较少, 不能充分覆盖参考评论中的关键信息. 这种模型虽然能生成结构合理的评论, 但在内容丰富性和细节上往往显得有所欠缺.

相比之下, BLIP-2 和本文的模型都显示出了对参考文献信息的全面理解和传达. 特别是, BLIP-2 模型虽然描述了洋葱圈的美味, 但其情感表达与参考文献中的情绪不完全匹配. 而本文的模型则在这方面做得更好, 不仅准确捕捉了洋葱圈“稍微干燥”的细节, 也与参考评论中的观点保持了高度一致, 展现了其在情感表达上的精准度. 总的来说, 本文的模型在生成评论时不只是信息全面, 还能精准表达情感. 这使得本文的模型更能满足用户的需求, 提高了用户体验.

4 总 结

本文提出了一项名为“多模态定制化评论生成”的新任务, 旨在基于产品图像和用户偏好生成定制化评论. 为此, 本文提出了一种基于预训练语言模型的多模态评论生成框架. 该框架为了在文本和图像模态之间建立无缝连接, 引入了一个图像标题生成组件. 此外, 在框架中还设计了一个由文本驱动的融合模块. 该模块进一步优化了不同模态间的交互, 确保生成的评论在语义上的连贯性和情感上的一致性. 实验结果充分验证了本文模型的有效性. 与传统的基准模型相比, 本文模型不仅提升了评论的信息量和准确性, 而且这些评论更能满足用户的多元化需求.

References

[1] Qiu G, Liu B, Bu JJ, Chen C. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 2011, 37(1): 9–27. [doi: 10.1162/coli_a_00034]

[2] Chen Z, Qian TY. Enhancing aspect term extraction with soft prototypes. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 2107–2117. [doi: 10.18653/v1/2020.emnlp-main.164]

[3] Bao XY, Wang ZQ, Jiang XT, Xiao R, Li SS. Aspect-based sentiment analysis with opinion tree generation. In: Proc. of the 31st Int'l Joint Conf. on Artificial Intelligence. Vienna, 2022. 4044–4050. [doi: 10.24963/ijcai.2022/561]

[4] Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. In: Proc. of the 10th Int'l Conf. on World Wide Web. Hong Kong: ACM, 2001. 285–295. [doi: 10.1145/371920.372071]

[5] He XN, Deng K, Wang X, Li Y, Zhang YD, Wang M. LightGCN: Simplifying and powering graph convolution network for recommendation. In: Proc. of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2020. 639–648. [doi: 10.1145/3397271.3401063]

[6] Tian CX, Xie YX, Li YL, Yang N, Zhao WX. Learning to denoise unreliable interactions for graph collaborative filtering. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Madrid: ACM, 2022. 122–132. [doi: 10.1145/3477495.3531889]

[7] Lu JS, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 13–23.

[8] Li G, Duan N, Fang YJ, Gong M, Jiang DX. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 11336–11344. [doi: 10.1609/aaai.v34i07.6795]

[9] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou YQ, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text Transformer. Journal of Machine Learning Research, 2020, 21(1): 5485–5551.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2021.

[11] Li JN, Li DX, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. of the 40th Int'l Conf. on Machine Learning. 2023. 19730–19742.

[12] Chen J, Wang SY, Zhao S, Zhang YP, Yu JY. Multi-granular user preferences for document-level sentiment analysis. Journal of Chinese Information Processing, 2023, 37(7): 122–130 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2023.07.015]

[13] Xia HL, Yang LS, Xue F. Large-scale Web document sentiment analysis method using self attention mechanism. Computer Engineering and Design, 2021, 42(9): 2642–2648 (in Chinese with English abstract). [doi: 10.16208/j.issn1000-7024.2021.09.032]

[14] Li WJ, Qi F, Yu ZT. Sentiment classification method based on multi-channel features and self-attention. Ruan Jian Xue Bao/Journal of Software, 2021, 32(9): 2783–2800 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5992.htm [doi: 10.13328/j.cnki.jos.005992]

[15] Nguyen HT, Le Nguyen M. Effective attention networks for aspect-level sentiment classification. In: Proc. of the 10th Int'l Conf. on Knowledge and Systems Engineering (KSE). Ho Chi Minh City: IEEE, 2018. 25–30. [doi: 10.1109/KSE.2018.8573324]

[16] Tulkens S, Van Cranenburgh A. Embarrassingly simple unsupervised aspect extraction. arXiv:2004.13580, 2020.

[17] Shi T, Li LQ, Wang P, Reddy CK. A simple and effective self-supervised contrastive learning framework for aspect detection. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 13815–13824. [doi: 10.1609/aaai.v35i15.17628]

[18] Wu ZX, Ong DC. Context-guided BERT for targeted aspect-based sentiment analysis. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 14094–14102. [doi: 10.1609/aaai.v35i16.17659]

[19] Gao L, Wang YL, Liu TC, Wang JY, Zhang L, Liao JX. Question-driven span labeling model for aspect-opinion pair extraction. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. 2021. AAAI Press, 12875–12883. [doi: 10.1609/aaai.v35i14.17523]

[20] Li JJ, Yu JF, Xia R. Generative cross-domain data augmentation for aspect and opinion co-extraction. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022. 4219–4229. [doi: 10.18653/v1/2022.naacl-main.312]

[21] Cai HJ, Tu YF, Zhou XS, Yu JF, Xia R. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 833–843. [doi: 10.18653/v1/2020.coling-main.72]

[22] Bu JH, Ren L, Zheng S, Yang Y, Wang JG, Zhang FZ, Wu W. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. arXiv:2103.06605, 2021.

[23] Zhao SM, Chen W, Wang TJ. Learning cooperative interactions for multi-overlap aspect sentiment triplet extraction. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi: Association for Computational Linguistics, 2022. 3337–3347. [doi: 10.18653/v1/2022.findings-emnlp.243]

[24] Gou ZB, Guo QY, Yang YJ. MvP: Multi-view prompting improves aspect sentiment tuple prediction. arXiv:2305.12627, 2023.

[25] Huang L, Lin CJ, He J, Liu HY, Du XY. Diversified mobile App recommendation combining topic model and collaborative filtering.

Ruan Jian Xue Bao/Journal of Software, 2017, 28(3): 708–720 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5163.htm [doi: 10.13328/j.cnki.jos.005163]

[26] Chen BY, Huang L, Wang CD, Jing LP. Explicit and implicit feedback based collaborative filtering algorithm. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 794–805 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5897.htm [doi: 10.13328/j.cnki.jos.005897]

[27] Choi J, Hong S, Park N, Cho SB. Blurring-sharpening process models for collaborative filtering. In: Proc. of the 46th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Taipei: ACM, 2023. 1096–1106. [doi: 10.1145/3539618.3591645]

[28] Sun PJ, Wu L, Zhang K, Su Y, Wang M. An unsupervised aspect-aware recommendation model with explanation text generation. ACM Trans. on Information Systems (TOIS), 2021, 40(3): 63. [doi: 10.1145/3483611]

[29] Shuai J, Zhang K, Wu L, Sun PJ, Hong RC, Wang M, Li Y. A review-aware graph contrastive learning framework for recommendation. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Madrid: ACM, 2022. 1283–1293. [doi: 10.1145/3477495.3531927]

[30] Ni JM, McAuley J. Personalized review generation by expanding phrases and attending on aspect-aware representations. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics,Vol. 2 (Short Papers). Melbourne: Association for Computational Linguistics, 2018. 706–711. [doi: 10.18653/v1/P18-2112]

[31] Li P, Tuzhilin A. Towards controllable and personalized review generation. arXiv:1910.03506, 2020.

[32] Sun PJ, Wu L, Zhang K, Fu YJ, Hong RC, Wang M. Dual learning for explainable recommendation: Towards unifying user preference prediction and review generation. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 837–847. [doi: 10.1145/3366423.3380164]

[33] Xi WD, Huang L, Wang CD, Zheng YY, Lai JH. Deep rating and review neural network for item recommendation. IEEE Trans. on Neural Networks and Learning Systems, 2022, 33(11): 6726–6736. [doi: 10.1109/TNNLS.2021.3083264]

[34] Ni JM, Lipton ZC, Vikram S, McAuley J. Estimating reactions and recommending products with generative models of reviews. In: Proc. of the 8th Int'l Joint Conf. on Natural Language Processing, Vol. 1 (Long Papers). Taipei: Asian Federation of Natural Language Processing, 2017. 783–791.

[35] Tang J, Yang YF, Carton S, Zhang M, Mei QZ. Context-aware natural language generation with recurrent neural networks. arXiv:1611.09900, 2016.

[36] Peng QY, Liu HT, Xu HY, Yang Q, Shao ML, Wang WJ. Review-LLM: Harnessing large language models for personalized review generation. arXiv:2407.07487, 2024.

[37] Atrey PK, Hossain MA, El Saddik A, Kankanhalli MS. Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 2010, 16(6): 345–379. [doi: 10.1007/s00530-010-0182-0]

[38] Bramon R, Boada I, Bardera A, Rodriguez J, Feixas M, Puig J, Sbert M. Multimodal data fusion based on mutual information. IEEE Trans. on Visualization and Computer Graphics, 2012, 18(9): 1574–1587. [doi: 10.1109/TVCG.2011.280]

[39] Zadeh A, Chen MH, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250, 2017.

[40] Tsai YHH, Bai SJ, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal Transformer for unaligned multimodal language sequences. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 6558–6569. [doi: 10.18653/v1/P19-1656]

[41] Huang J, Tao JH, Liu B, Lian Z, Niu MY. Multimodal Transformer fusion for continuous emotion recognition. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP 2020). Barcelona: IEEE, 2020. 3507–3511. [doi: 10.1109/ICASSP40776.2020.9053762]

[42] Yang HZ, Gao XQ, Wu JL, Gan T, Ding N, Jiang FJ, Nie LQ. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In: Findings of the Association for Computational Linguistics: ACL 2023. Toronto: Association for Computational Linguistics, 2023. 6267–6281. [doi: 10.18653/v1/2023.findings-acl.390]

[43] Zhu JN, Li HR, Liu TS, Zhou Y, Zhang JJ, Zong CQ. MSMO: Multimodal summarization with multimodal output. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 4154–4164. [doi: 10.18653/v1/D18-1448]

[44] Chung HW, Hou L, Longpre S, *et al*. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 2024, 25(1): 3381–3433.

[45] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision (ECCV 2014). Zurich: Springer Int'l Publishing, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[46] Yan A, He ZK, Li JC, Zhang TY, McAuley J. Personalized showcases: Generating multi-modal explanations for recommendations. In: Proc. of the 46th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Taipei: ACM, 2023. 2251–2255. [doi:

10.1145/3539618.3592036]
[47] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2017.
[48] Paulus R, Xiong CM, Socher R. A deep reinforced model for abstractive summarization. arXiv:1705.04304, 2017.
[49] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.
[50] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 311–318. [doi: 10.3115/1073083.1073135]
[51] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proc. of the 9th Workshop on Statistical Machine Translation. Baltimore: Association for Computational Linguistics, 2014. 376–380. [doi: 10.3115/v1/W14-3348]
[52] Lewis M, Liu YH, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461, 2019.
[53] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, 2023.
[54] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 27730–27744.
[55] Wang JF, Yang ZY, Hu XW, Li LJ, Lin K, Gan Z, Liu ZC, Liu C, Wang LJ. GIT: A generative image-to-text Transformer for vision and language. arXiv:2205.14100, 2022.
[56] Lee K, Joshi M, Turc I, Hu HX, Liu FY, Eisenschlos J, Khandelwal U, Shaw P, Chang MW, Toutanova K. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In: Proc. of the 40th Int'l Conf. on Machine Learning. 2023. 18893–18912.
[57] Liu HT, Li CY, Wu QY, Lee YJ. Visual instruction tuning. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 34892–34916.
[58] Li B, Lv CH, Zhou ZF, Zhou T, Xiao T, Ma AX, Zhu JB. On vision features in multimodal machine translation. arXiv:2203.09173, 2022.
[59] Ling Y, Yu JF, Xia R. Vision-language pre-training for multimodal aspect-based sentiment analysis. arXiv:2204.07955, 2022.
[60] Zhou R, Guo WY, Liu XM, Yu SL, Zhang Y, Yuan XJ. AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis. arXiv:2306.01004, 2023.

附中文参考文献

[12] 陈洁, 王思雨, 赵姝, 张燕平, 余静莹. 基于多粒度用户偏好的文档级情感分析. 中文信息学报, 2023, 37(7): 122–130. [doi: 10.3969/j.issn.1003-0077.2023.07.015]
[13] 夏辉丽, 杨立身, 薛峰. 利用自注意力机制的大规模网络文档情感分析. 计算机工程与设计, 2021, 42(9): 2642–2648. [doi: 10.16208/j.issn1000-7024.2021.09.032]
[14] 李卫疆, 漆芳, 余正涛. 基于多通道特征和自注意力的情感分类方法. 软件学报, 2021, 32(9): 2783–2800. http://www.jos.org.cn/1000-9825/5992.htm [doi: 10.13328/j.cnki.jos.005992]
[25] 黄璐, 林川杰, 何军, 刘红岩, 杜小勇. 融合主题模型和协同过滤的多样化移动应用推荐. 软件学报, 2017, 28(3): 708–720. http://www.jos.org.cn/1000-9825/5163.htm [doi: 10.13328/j.cnki.jos.005163]
[26] 陈碧毅, 黄玲, 王昌栋, 景丽萍. 融合显式反馈与隐式反馈的协同过滤推荐算法. 软件学报, 2020, 31(3): 794–805. http://www.jos.org.cn/1000-9825/5897.htm [doi: 10.13328/j.cnki.jos.005897]

作者简介

强敏杰, 博士生, CCF 学生会员, 主要研究领域为自然语言处理.
王中卿, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理.
李寿山, 博士, 教授, CCF 专业会员, 主要研究领域为自然语言处理.
周国栋, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理.